Histogram-constrained Image Generation

Haoming Liu[1,2], Yuanhe Guo[1,2], Yijia Cao[1], Shenji Wan[1,2], and Hongyi Wen[1,2]

[1] Center for Data Science, New York University Shanghai, Shanghai, China
[2] New York University, New York, USA
{haoming.liu, hongyi.wen}@nyu.edu

Abstract. Diffusion models have emerged as a dominant paradigm in generative modeling, enabling high-fidelity sampling from complex data distributions. Despite impressive capabilities, controlling diffusion models to produce outputs aligned with user intent remains an open challenge, especially when balancing global coherence with local precision. Existing control mechanisms vary in the granularity of their conditioning signals. For example, textual prompts guide generation globally through high-level semantics, while ControlNet-like approaches secure precise local structure via dense conditions. In this work, we introduce **H**istogram-constrained **I**mage **G**eneration (**HIG**), a novel control mechanism that falls into the middle ground of control granularity. Our framework enforces user-specified distributional constraints (e.g., color histograms or latent token distributions) during the generation process with exact precision. We model such control as an optimal transport (OT) problem and apply explicit guidance transformations during sampling, thereby driving the diffusion trajectory to align with the desired histogram. We demonstrate the versatility of HIG across diverse applications, including constrained generation via color/latent histograms and high-capacity information embedding through histogram-level encoding. Our findings underscore the promise of distributional control, a flexible and interpretable control scheme that is fully compatible with existing control mechanisms, diversifying the hybrid strategies for controllable image generation. Our project page is available at: https://maps-research.github.io/hig/.

Keywords: Diffusion Models · Constrained Generation

1 Introduction

Diffusion models have emerged as a dominant paradigm in generative modeling, enabling high-fidelity sampling from complex data distributions [9,20,55,56]. As their capabilities grow, there has been an increasing focus on controllable generation, guiding the sampling process to align with user intent. Extensive work explores this direction across diverse control schemes, including text prompts [27, 46,50], structural guidance [32,73], and content/style personalization [21,51].

Intuitively, controllable generation methods can be sorted along a spectrum defined by the information granularity of their control signals. On one end, high-level control schemes guide generation through coarse-grained signals, such as text prompts [12,27,46] or content/style LoRAs [21,51]. These signals typically

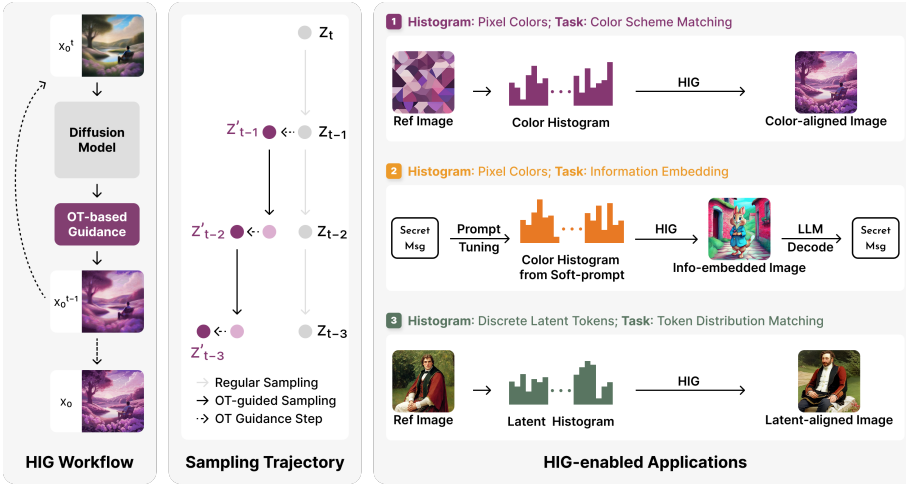

Fig. 1: Overview for HIG. We intervene in the diffusion process with explicit OT-based guidance. HIG enables diverse applications, including constrained generation with arbitrary histogram constraints and high-capacity information embedding.

encode abstract concepts and grant the diffusion process considerable flexibility to improvise during generation. As a result, they influence the output at a global scale, such as specifying the subjects or artistic styles. On the other end, low-level control schemes like ControlNet [32,73] typically condition on dense inputs such as edge maps, depth maps, or pose annotations. These methods provide fine-grained spatial control and are particularly effective at anchoring the generation to a given structure or layout.

In this work, we explore a middle ground on the spectrum. We introduce **H**istogram-constrained **I**mage **G**eneration (**HIG**), a novel control scheme that regulates the distributional properties of generated images. HIG is characterized by three key properties:

1. **Flexibility**: it can enforce distributional consistency across diverse choices, such as histograms of pixel color or latent token, enabling wide applications.
2. **Precision and Optimality**: the constraint is formalized as an optimal transport (OT) problem [61], which identifies the minimal-cost transformation required to satisfy the constraint exactly (Sec. 3.2). This optimality enables implementing control through a direct intervention that perturbs the intermediate diffusion outputs. As the perturbation is guaranteed to be minimal in cost, the guidance is mathematically principled and minimally invasive. In addition, the ability to satisfy the constraint precisely opens up applications that demand exact precision (e.g., information embedding).
3. **Lightweight and Compatible Design**: HIG is training-free, and its inference-time overhead is negligible compared to the diffusion generation itself. Unlike most existing control mechanisms that rely on auxiliary inputs or modified architectures, HIG is plug-and-play, interpretable, and compatible with many established methods, such as ControlNets [32,73] and LoRAs [21,51].

HIG unlocks a wide range of applications. First, HIG enables generating images that exactly match a given color histogram, offering controls on the global property (overall color schemes) with precision (Sec. 4.1). Second, we demonstrate the control precision of histogram constraints through an information embedding technique, where we embed information via the histogram so that the generated image entails the same information. For example, by deterministically mapping a soft-prompt embedding [31] (i.e., a continuous-valued vector) to a target color histogram, our method can encode hundreds of text tokens within a visually indistinguishable image, which can be precisely decoded based on the color histogram (Sec. 4.2). Lastly, we showcase HIG's generalization to latent histograms, where the histograms are computed from the discrete codebooks of image tokenizers [13,48,49,60,71]. In such cases, the control effect of latent histograms hinges on the information captured in the latent space, such as low-level colors and textures or high-level semantics. We demonstrate the versatility and control precision of HIG through detailed experiments and analysis.

2 Inference-Time Guidance via Explicit Transformations

Diffusion models are a class of generative models that learn to map a simple prior distribution (typically Gaussian) to a complex data distribution through a series of time-dependent interpolations. It involves a forward process that progressively corrupts the data, and a reverse process that recovers data from prior. Various mathematical formulations have been proposed to characterize this process, including discrete-time probabilistic models [20], continuous-time score SDEs [57], and flow matching [35]. Here, we follow the DDPM/DDIM [20,56] formulation to formally describe the proposed inference-time guidance.

The predominant paradigm for high-resolution image synthesis is to train latent diffusion models (LDM) that operate in the latent space of a VAE [26] and perform conditional generation in text-to-image fashion [46,50]. At each sampling step, the LDM takes in the noisy latent $\mathbf{z}_t$ and text condition $\mathbf{c}$ and predicts the quantities that correspond to the training objective (e.g., noise, signal, or velocity), which essentially characterizes the current view of the final image latent $\mathbf{z}_0$. Then, a controlled amount of noise is added back to the intermediate guess $\mathbf{z}_0^t$ to proceed to the less noisy latent $\mathbf{z}_{t-1}$ at the next time step. Formally, at time step t, the intermediate guess of a noise-prediction based LDM can be expressed as follows:

$$\mathbf{z}_0^t = \frac{1}{\sqrt{\bar{\alpha}_t}}\mathbf{z}_t - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t), \tag{1}$$

where $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)$ denotes the noise prediction by model θ and $\{\bar{\alpha}\}_{t=1:T}$ denote the cumulative noise factors.

Notably, most existing control mechanisms are achieved by manipulating $\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t)$ via auxiliary inputs (ControlNets) or model pieces (LoRAs), thereby deviating from the original trajectory. In contrast, we adopt a complementary approach that explicitly transforms the intermediate prediction $\mathbf{z}_0^t$. Formally,

given a guidance transformation φ that operates on images, the transformed latent $\mathbf{z}_0^{t'}$ can be obtained through a decode-transform-encode cycle:

$$\mathbf{z}_0^{t'} = \text{VAE-encode}(\varphi(\text{VAE-decode}(\mathbf{z}_0^t))). \tag{2}$$

Hence, we can step to the less noisy latent $\mathbf{z}_{t-1}$ based on the transformed latent:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{z}_0^{t'} + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{z}_t, \mathbf{c}, t). \tag{3}$$

Typically, such perturbation only needs to be applied to a subset of the sampling time steps, denoted as $\mathcal{T}$. We note that the OT formulation ensures the intervention is minimally invasive by explicitly minimizing transportation cost. As a result, the transformation remains close to the original diffusion trajectory, allowing high-fidelity images to be generated. In this way, the proposed guidance scheme enables explicit control over the generated image during de-noising, while allowing the subsequent diffusion steps to further refine the final output.

3 Image Generation with Distributional Guidance

3.1 Preliminaries

We define a histogram as a discrete probability distribution over a finite set of d bins, denoted as a vector $\mathbf{h} \in \mathbb{R}^d$ with non-negative entries summing to one. For data with continuous values, such as RGB colors in $[0, 1]^3$, we partition the space into d disjoint bins by uniformly dividing each dimension. For example, a bin may correspond to the colors in $[0, 0.1]^3$, and its representative color can be defined as the mean of each sub-interval. For discrete domains such as VQ-VAE tokens [60], bins naturally correspond to individual token indices, and we can simply count the per-token frequency. Notably, a bin may also correspond to an arbitrary set of tokens/items (e.g., a palette of colors, or a certain subset of VQ codebook tokens), allowing flexible and task-specific binning strategies. For clarity, we term the binning strategy that maps to a single token/item as *single-option binning* and to a subset of tokens/items as *multi-option binning*.

3.2 Histogram Matching with Optimal Transport

Given the source image $\mathcal{I}$, we wish to solve for a transformed image $\mathcal{I}' = \varphi(\mathcal{I}, \mathbf{h}^{tgt})$ with minimal content deviation, where φ denotes a guidance transformation that matches $\mathcal{I}$ to the target histogram $\mathbf{h}^{tgt}$. To achieve this, we employ optimal transport (OT) [61], a mathematically grounded approach that identifies the most efficient way to transform one distribution into another. In addition to the minimal transformation cost, OT guarantees exact compliance with the target histogram by producing a transport plan that explicitly specifies the quantity of individual units (i.e., number of pixels or tokens) to be reassigned between each source and target bin.

We first discuss the OT formulation for single-option binning, where each histogram bin only consists of a single color range or discrete token. Formally,

given the source histogram $\mathbf{h}^{src}$ and the target histogram $\mathbf{h}^{tgt}$, with $\mathbf{h}^{src}, \mathbf{h}^{tgt} \in \mathbb{R}^d$, we want to solve the following optimization problem:

$$\gamma = \arg\min_{\gamma} \langle \gamma, \mathbf{M} \rangle_F, \quad \text{s.t. } \gamma \mathbf{1} = \mathbf{h}^{src}, \gamma^\top \mathbf{1} = \mathbf{h}^{tgt}, \gamma \geq 0, \tag{4}$$

where $\gamma \in \mathbb{R}^{d \times d}$ denotes the optimal transport plan, $\mathbf{M} \in \mathbb{R}^{d \times d}$ denotes the cost matrix. The cost matrix can be pre-computed based on some distance metric between color tuples or latent embeddings. The OT plan γ informs the proportion of every source bin to be transported to a target bin to achieve the desired histogram. We implement the OT transformation by randomly sampling the source pixels/tokens and setting them to the target bin values.

In some cases, strict single-option binning may lead to excessive content distortion during OT-based histogram matching. To mitigate this, we introduce a multi-option binning scheme for OT, where each bin contains multiple candidate values. In this setting, the transport plan only enforces the aggregated mass per bin to match the target histogram, allowing flexible assignments among the intra-bin options to reduce perceptual deviation.

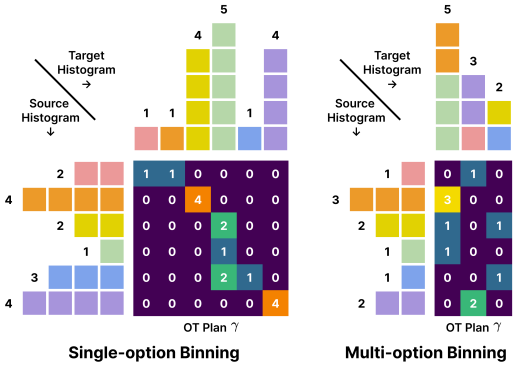

Fig. 2: Exemplar OT plans with single-option ($d = 6$) and multi-option binning ($k = 2, d = 3$).

Unlike the single-option setting, the multi-option case requires a more sophisticated formulation due to the added ambiguity within each bin. Formally, let each of the d target bins contain k candidate values, yielding a total of kd options. The OT plan is now defined over a $kd \times d$ matrix, where each row corresponds to a specific source option (e.g., a pixel or token value), and each column denotes a target bin. In this case, the target distribution constraint is applied only at the bin level. Within each bin, however, the assignment to specific options remains unconstrained, allowing flexibility in choosing the most perceptually similar mappings. We provide binning illustrations in Fig. 2.

To construct the corresponding cost matrix, we first consider the pairwise distances between all possible options, yielding a cost matrix of size $kd \times kd$. For each entry in the $kd \times d$ OT plan, the cost is defined as the minimum distance from the source option to any of the k options in the target bin. This ensures that, for any source option and target bin, the transport plan always transforms it to the "closest match" in the target bin, reducing unnecessary distortion while still satisfying the bin-level constraints. Formally, we consider $\mathbf{h}^{src} \in \mathbb{R}^{kd}$ and $\mathbf{h}^{tgt} \in \mathbb{R}^d$ in multi-option binning. For the j-th option in the i-th source bin, its

closest match in the p-th target bin can be pre-computed and assigned to the cost matrix $\mathbf{M} \in \mathbb{R}^{kd \times d}$, such that:

$$\mathbf{M}_{ik+j,p} = \min_{q} \text{dist}(\mathbf{v}_{ik+j}, \mathbf{v}_{pk+q}), \tag{5}$$

where $i, p \in \{0, 1, \cdots, d-1\}$, $j, q \in \{0, 1, \cdots, k-1\}$, $\mathbf{v}$ corresponds to the stacked option vectors (e.g., kd-many color tuples or latent embeddings), $\text{dist}(\cdot)$ denotes a distance metric, such as L1 or L2 distance. When applying the transport plan, we randomly sample a given number of individual units from each source bin and set them to the closest option in the target bin, respectively.

In practice, we use the vanilla network simplex algorithm [1] to solve for the OT plans. For color histogram matching on 1024^2 images with $d = 4096$, the optimization takes roughly 0.2 seconds (maxIter = 500K). While faster approximate OT solvers such as Sinkhorn [8] (with entropic regularization) are available, the vanilla solver is already good enough for our usage.

4 Construction of Target Histograms

4.1 Histograms from Reference Images

As a direct approach, we can use histograms extracted from reference images as the target. Specifically, we calculate the histogram of pixel colors or latent tokens and set it as $\mathbf{h}^{tgt}$. During sampling, we apply the guidance transformation φ by transporting pixels or tokens according to an optimal transport (OT) plan, resulting in a perturbed output that precisely matches the target histogram.

Notably, we can flexibly adjust the control precision depending on the usage. For color histograms, one may optionally perform a post-hoc OT step on the generated image to enforce exact compliance with $\mathbf{h}^{tgt}$. This guarantees histogram alignment but may introduce visual artifacts (e.g., locally inconsistent pixels), due to the rigid reassignment of colors. Such exact control is necessary in use cases that demand perfect compliance to the histogram constraint (e.g., information embedding), but it can be safely omitted in more relaxed settings (e.g., color scheme matching for aesthetics).

4.2 Histograms from Continuous Vectors

In addition to using reference images, we can also construct target histograms from continuous vectors. We showcase this with an *information embedding* technique, whose workflow is illustrated in Fig. 3. Intuitively, the objective is to embed hidden information into an image such that it can be reliably recovered through decoding. In practice, we extend our color-constrained generation framework by replacing the target histogram with a synthetic color distribution derived from a continuous-valued vector (referred to as a soft-prompt embedding). The soft-prompt embedding is optimized to condition an LLM to reproduce a target sentence. Since accurate decoding requires exact alignment between the color histogram and the embedding, this task highlights the strength of our method in enforcing precise constraints through explicit OT-based guidance.

We first elaborate on how a sequence of text tokens can be transformed into a compact soft-prompt embedding via prompt tuning [31]. Prompt tuning is a parameter-efficient fine-tuning (PEFT) technique that learns a set of continuous embeddings (soft prompts) that are prepended to the input text to guide the language model's behavior, which can be viewed as task-specific instructions in the embedding space. In our case, we need a soft prompt that can condition the LLM to output an exact sequence of text tokens. Formally, we aim to maximize the conditional log-likelihood of the target text sequence $\mathbf{y} = (y_1, y_2, \cdots, y_T)$ given the prepended soft prompt $\mathbf{p} \in \mathbb{R}^d$. With a decoder-only LLM parameterized by Θ, the optimization objective can be expressed as:

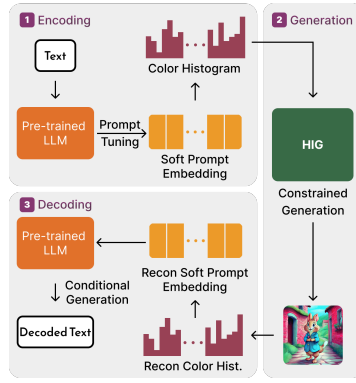

Fig. 3: Illustration of our information embedding workflow.

$$\max_{\mathbf{p},||\mathbf{p}||=B} \mathcal{L}(\mathbf{p}) = \max_{\mathbf{p},||\mathbf{p}||=B} \sum_{t=1}^{T} \log P(y_t|\mathbf{p}, y_{<t}; \Theta), \tag{6}$$

where $P(y_t|\mathbf{p}, y_{<t}; \Theta)$ denotes the probability of the token y_t at time step t, conditioned on the soft prompt $\mathbf{p}$ and the previously generated tokens $y_{<t}$. Though more advanced algorithms could be applied, we find that the simple gradient ascent suffices to optimize Eq. 6, and the norm constraint is enforced by setting every iteration's result to a fixed norm B (for simpler inverse mapping, discussed below). In addition, we empirically observe that tuning a single-token soft-prompt vector is sufficient to condition the selected LLM (Llama-3.1-8B [11]) to output the exact sequence up to hundreds of tokens (within fewer than 1K optimization steps). After the optimization stage, we can then transform the soft-prompt embedding $\mathbf{p}$ to the target color distribution $\mathbf{h}^{tgt} \in \mathbb{R}^d$ through a simple deterministic mapping: $\mathbf{h}^{tgt} = f(\mathbf{p}) = \frac{\exp(\mathbf{p})}{\sum_i \exp(\mathbf{p}_i)}$, where the exponential function ensures the non-negativity and the normalization ensures all the entries of $\mathbf{h}^{tgt}$ sum to 1. As for the inverse mapping f^{-1}, we are essentially looking for some scaling factor $k \in \mathbb{R}$ such that $||\ln(k\mathbf{h}^{tgt})|| = ||\mathbf{p}||$. Given the fixed norm B, the value of k can be efficiently determined, and under a fixed branch convention we recover a unique $\mathbf{p}$, establishing a one-to-one mapping between $\mathbf{p}$ and $\mathbf{h}^{tgt}$, where no additional information is required for decoding.

5 Experiment

5.1 Implementation Details

Histogram Configuration. For color-constrained generation with reference images, we use $d = 4096$ and single-option binning, where the color channels

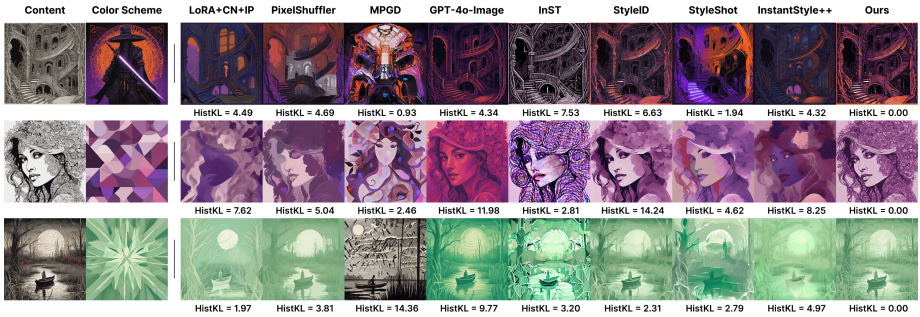

Fig. 4: Qualitative results on color-constrained generation. "LoRA+CN+IP" refers to the stacked control from LoRA [51], ControlNet [32], and IP-Adapter [69]. HistKL quantifies the KL divergence to the target color distribution (the lower the better).

are quantized to match the histogram dimension (e.g., 16^3 for RGB binning, 64^2 for RG binning, etc.). For information embedding, we employ Llama-3.1-8B [11] for soft prompt tuning, with $d = 4096$, fixed norm $B = 40.0$. We use the AdamW [39] optimizer with step-based learning rate decay. Empirically, the soft prompt for most text sequences of fewer than 500 tokens converges within 500 optimization steps. For multi-option binning, we use $k = 16$ and $d = 4096$, resulting in 65536 color options. All cost matrices are pre-computed based on the L1 distance. We adopt the OT solvers offered by Python Optimal Transport (POT) [3]. For latent-constrained generation, we keep the histogram dimension aligned with the codebook size of image tokenizers.

Image Generation. To compare with relevant baselines fairly, we use `SDXL` [46] as the base model for image generation (on 1024^2 resolution), with a DDIM [56] noise scheduler and 50 sampling steps. Meanwhile, HIG generalizes to newer models trained with flow-based objectives [35, 37]. The guidance time steps $\mathcal{T}$ are set by the usage. For histograms derived from reference images, 1-2 OT-based transformations are sufficient to offer decent distributional guidance. For the synthetic histograms derived from the soft-prompt embeddings, we generally need 3-4 OT-based transformations to get better generation quality. We direct the readers to the Appendix for more details on the generalization to `FLUX.1[dev]` [27] and the ablation studies on the optimal choice of $\mathcal{T}$.

5.2 HIG for Color-constrained Generation

We compare the color-constrained generation capability of HIG with other relevant baselines, including GPT-4o-Image [42], stacked controls using DreamBooth LoRA [51], ControlNet++ [32], IP-Adapter [69], soft-shuffling guided color scheme matching [72], manifold preserving guidance [18], and style transfer approaches [7, 15, 62, 74]. The curated benchmark consists of 300 textual prompts (sourced from Civitai [4]) and 300 reference images (half depicting certain content

[3] https://pythonot.github.io/
[4] https://civitai.com/

Method	Primary Task	HistKL ↓	CLIP ↑	Aesthetics ↑	Training-free?
Unconstrained	Text-to-Image Generation	10.87	29.47	6.98	-
LoRA+CN+IP	Color/Structure/Style Control	3.16	22.53	5.39	✗
PixelShuffler	Color Scheme Matching	5.90	23.56	4.87	✗
MPGD	Constrained Generation	4.24	24.46	4.94	✓
GPT-4o-Image	In-Context Generation	8.96	25.72	6.62	✓
InST	Style Transfer	5.67	17.41	4.83	✗
StyleID	Style Transfer	11.32	24.76	6.18	✓
StyleShot	Style Transfer	1.90	26.39	5.83	✓
InstantStyle++	Style Transfer	2.72	26.23	6.54	✓
Ours (direct OT)	Img-to-Img Translation	**0.00**	26.94	6.52	✓
Ours (guidance)	Constrained Generation	1.09	**27.19**	**6.78**	✓
Ours (guidance + post-hoc OT)	Constrained Generation	**0.00**	26.91	6.66	✓

Table 1: Quantitative results for color-constrained generation. The best results are shown in **bold**. The "direct OT" variant applies OT-based transformation to the unconstrained images; whereas the "guidance" variants apply OT-based guidance transformations during the sampling process.

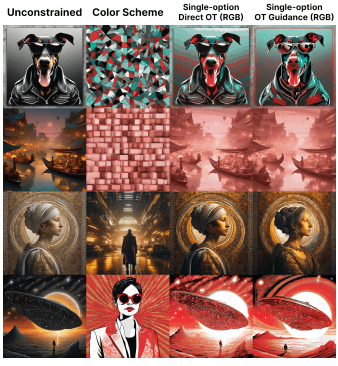

Fig. 5: Qualitative results for color-constrained image generation. OT-based guidance helps alleviate visual artifacts.

Method	Base Model	Latency (s) ↓	Overhead (s) ↓
Unconstrained	SDXL	10.67	–
HIG (w/o post-hoc OT)	SDXL	12.87	2.20
HIG (w/ post-hoc OT)	SDXL	15.06	4.39
DreamBooth LoRA*	SDXL	13.01	2.34
ControlNet++ (Depth)	SDXL	25.47	14.80
ControlNet++ (Softedge)	SDXL	15.51	4.84
ControlNet++ (OpenPose)	SDXL	17.19	6.52
InstantStyle++	SDXL	79.52	68.85
InST*	SD1.4	4.88	N/A
MPGD	SD1.4	10.83	N/A
StyleID	SD1.4	23.89	N/A
StyleShot	SD1.5	44.91	N/A
PixelShuffler*	N/A	38.42	N/A

Table 2: Latency of different control baselines. By default, we use 50 sampling steps and FP16 precision on a single NVIDIA A100 GPU. HIG's overhead is on par with ControlNets and LoRAs. * indicates that the method requires sample-specific tuning.

and half composed of abstract color schemes). Our evaluation includes both qualitative (Fig. 4) and quantitative (Tab. 1) aspects, measuring histogram alignment (HistKL), prompt compliance (CLIP-Score [19]), and visual aesthetics (LAION-Aesthetics [52]). Overall, the results show that HIG consistently outperforms other baselines in terms of histogram alignment, prompt compliance, and visual quality, regardless of whether post-hoc OT is applied. In addition, we observe that applying OT-based guidance transformations during the denoising process yields notable gains over the "direct OT" variant. As demonstrated in Fig. 5 (row 1), performing OT-based guidance during sampling helps alleviate the visual artifacts caused by the locally inconsistent pixels. Lastly, we evaluate the overheads incurred by different control baselines (Tab. 2). Overall, HIG incurs minimal computational overhead and is roughly comparable to common control methods like LoRAs and ControlNets.

Notably, HIG is not intended as a drop-in replacement for specialized tasks such as style transfer, where dedicated methods are specifically designed and highly optimized. Instead, our goal is to demonstrate that HIG provides strong

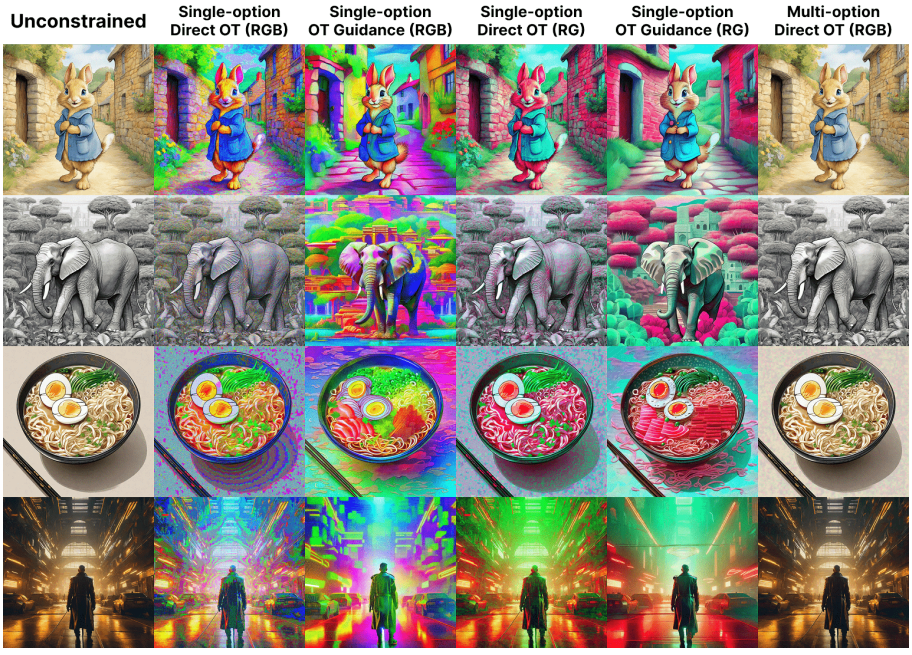

Fig. 6: Qualitative results of information embedding. Each image embeds 512 text tokens that can be faithfully decoded. Under single-option binning, OT-based guidance (col 3&5) drastically reduces visual artifacts compared to direct OT variants (col 2&4); under multi-option binning, the embedded images remain visually similar to unconstrained generations (col 6). Better view with colors.

and reliable control over color distributions while maintaining high visual fidelity and low computational overhead. Compared to relevant control baselines, HIG achieves a favorable balance between controllability, image quality, and efficiency, making it a practical solution for color-constrained generation scenarios.

5.3 HIG for Information Embedding

The evaluation of histogram-based information embedding consists of two stages: in the first stage, we perform Prompt Tuning [31] to optimize for the soft-prompt embedding that can condition the LLM to output the exact text, where we observe the success rate of finding desired soft-prompts; in the second stage, we convert the soft-prompt embeddings to color histograms and perform constrained generation under such guidance, where the objective is to monitor the rate of exact information decoding from the generated images.

To obtain complex text sequences to embed in images, we collect the Markdown code from popular GitHub repositories [5], providing a mixture of multilingual text, code, and URLs. We randomly crop the raw text to have a fixed number of tokens within {32, 64, 128, 256, 512}, resulting in 300 text sequences

[5] https://github.com/EvanLi/Github-Ranking

at each token length. For constrained generation, we adopt $\mathcal{T} = \{40, 30, 20, 10, 0\}$ (i.e., with post-hoc OT) to improve content fidelity and secure precise histogram alignment. By the qualitative results in Fig. 6, we observe that: 1) given the synthetic color histograms derived from soft prompt embeddings, performing direct OT on the unconstrained image tends to give coarse and noisy output under single-option binning (col 2&4); 2) applying OT-based guidance during the denoising process significantly improves the generation quality while ensuring histogram alignment (col 3&5); 3) direct OT with multi-option binning results in minor content deviations from the source image (col 6), which is courtesy of relaxing the color-level constraint to bin-level.

The quantitative results are shown in Tab. 3. In general, the optimization of a text sequence up to 256 text tokens takes less than 20 seconds with a >99.7% success rate (evaluated on a single NVIDIA A100 GPU). For the decoding side, the exact match rate of both binning strategies slightly decreases as we enlarge the text sequence length. In the case of 512 tokens, the rate of decoding the exact target text for both binning strategies remains above 90%. Note that it is empirically possible to further improve the exact decoding rate by enlarging the image resolution or using multiple soft-prompts. Meanwhile, image quality stays largely unaffected as the embedded text length grows. This is because the soft-prompt embedding is of fixed dimension and always maps to a color distribution of the same scale: image complexity is therefore decoupled from text length, while the exact decoding rate is bounded by how finely a finite pixel budget can approximate a continuous distribution, which improves as the image resolution grows.

Embedded Text Tokens	Soft Prompt Tuning		Binning Strategy	Decoding
	Success ↑	Time (s) ↓		Exact Match ↑
32	100.0%	4.87	Single	100.0%
			Multiple	100.0%
64	99.7%	6.19	Single	99.7%
			Multiple	99.7%
128	99.7%	11.16	Single	99.7%
			Multiple	99.7%
256	99.7%	20.06	Single	97.3%
			Multiple	97.7%
512	98.3%	300.08	Single	91.0%
			Multiple	90.3%

Table 3: Quantitative results on information embedding under single-option binning (with OT guidance) and multi-option binning (with direct OT). Both variants compute histograms from RGB channels.

As shown in Fig. 7, we further evaluate the robustness of our information embedding technique, where we adopt the HIG variant with single-option binning over RGB channels. Under random scaling (with scale factor uniformly sampled from $[0.5, 2.0]$), the exact reconstruction rate remains around 90% for text lengths up to 128 tokens. In contrast, the generated images are more sensitive to JPEG compression. The exact match rate drops below 90% once the embedded text exceeds 32 tokens. We also examine robustness to soft-prompt perturbation and histograms with wrongly binned values. With fewer than 5% wrongly binned values or Gaussian noise, the decoding success rate remains above 90% at a sequence length of 128 tokens. To sum up, our results indicate that the proposed information embedding technique is robust to common perturbations, with robustness increasing as the embedded sequence becomes shorter.

Overall, the information embedding results highlight a key advantage of HIG: its ability to enforce exact distributional control, which is essential for reliable

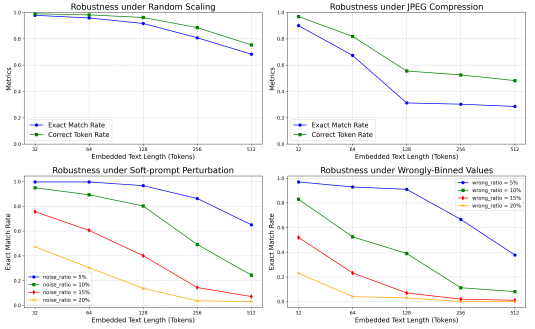

Fig. 7: Robustness evaluation of our information embedding technique. Our evaluation spans random scaling, JPEG compression, soft-prompt perturbation, and histogram corruption.

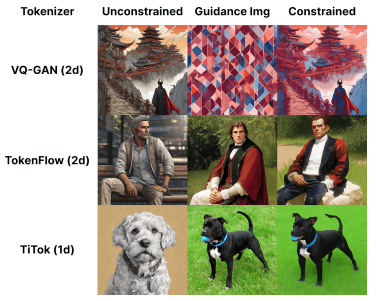

Fig. 8: Manipulating latent token histograms with different image tokenizers results in varying control effects (e.g., colors, semantics).

information decoding. Without precise histogram matching, the reconstructed soft prompt will deviate from the target statistics and decode into random text fragments rather than the intended message.

5.4 HIG for Latent Histogram Matching

To further examine the generalization capability of HIG, we extend our framework to latent histograms, where the goal is to perform constrained generation using target histograms formed by discrete latent tokens. The procedure for OT guidance on latent histograms is as follows: 1) tokenizing the guidance image and intermediate image prediction into latent token maps, 2) performing OT-based guidance transformation based on the token distributions derived from the guidance image, and 3) decoding the perturbed tokens to obtain the transformed output. We apply this workflow across multiple image tokenizers, including VQ-GAN [13], TokenFlow [48], and TiTok [71].

As illustrated in Fig. 8, the observed control effects are closely tied to the properties of the latent space. For tokenizers that capture low-level visual features (e.g., colors and textures), HIG produces effects akin to color scheme matching. In contrast, when operating in tokenizers that encode high-level semantics, manipulation of the latent histograms yields semantic-level guidance. Lastly, for 1D tokenizers, we observe strong control over spatial structures.

These observations suggest that different image tokenizers encode distinct levels of abstraction (e.g., colors, semantics, etc.) and exhibit varying degrees of robustness, which in turn lead to different control behaviors under latent histogram guidance. Although the OT guidance guarantees a theoretically optimal solution to transform token distributions, the perceptual quality and stability of the resulting perturbations depend critically on the robustness of the latent space and its decoder. In general, more robust latent representations allow finer and more controllable deviations, whereas less stable spaces may produce coarser or

even collapsed outputs. In the Appendix, we present further studies on enabling color scheme and semantic control with latent histogram matching.

As a proof-of-concept, our results on latent histograms show that applying HIG in latent space is feasible, and the control effects are inherently dependent on the information captured by the latent space as well as the robustness of the decoder. A systematic investigation of control behaviors across different latent spaces remains a promising direction for future work.

6 Related Work

Text-to-Image Diffusion Models. Following Rombach *et al.* [50], recent approaches adopt the LDM framework, with a trend of switching to the DiT-style architectures [43] and scaling up the models [46]. Many recent works have incorporated new techniques for model training, such as flow matching [35, 37] and EDM [23, 24], resulting in a series of competitive models [3, 12, 16, 27, 64]. We note that our framework is widely applicable to any text-to-image model. As long as it requires multiple steps during sampling, we can always perturb the intermediate predictions and resume the generation in the same spirit. We present additional results of applying HIG on flow-based text-to-image models (e.g., FLUX.1[dev] [27]) in the Appendix.

Controllable Generation. There have been a variety of works that aim to introduce additional control signals. LoRA-based approaches utilize parameter-efficient fine-tuning (PEFT) to learn adaptation parameters that specialize the content or style from reference images [21, 36, 51, 53]. ControlNet-like approaches [32, 40, 59, 73, 75] introduce extra parameters to enforce fine-grained control over local regions based on dense spatial signals. Image colorization [25, 34, 72] and style transfer [7, 15, 62, 74] approaches transfer the color scheme or artistic styles from reference images to other images. Some works take multimodal inputs and perform in-context image generation and editing [28, 42, 63]. Meanwhile, a recent line of work targets color control in diffusion models: SW-Guidance [38] steers sampling toward the color distribution of a reference image with a differentiable Sliced-Wasserstein objective, Color Alignment [54] trains a diffusion model to generate within a given color pattern, mapping intermediate samples to their nearest reference colors, and ModFlows [29] matches the color scheme with a separately trained, post-hoc transfer model that gives a flow-based approximation of the OT map. Other methods operate on more localized color cues, binding a target color to a specific object through cross-attention [30] or conditioning on a color palette [47]. HIG is complementary to these efforts and adopts a different design: it takes the full color histogram as the control target and enforces it with an explicit, minimal-cost transport plan, so the generated distribution matches the target *exactly*, training-free and without inference-time gradients.

Training-free Guidance. The OT-based guidance in HIG can be viewed as a form of training-free guidance (TFG). Most TFG approaches rely on heuristic-

based gradients [2,5,68], whereas HIG operates in a possibly non-differentiable way by inserting a user-defined transformation into the loop. There are also derivative-free techniques that optimize non-differentiable targets by sampling multiple at each step and selecting ones with the highest rewards [10,17,22,33].

Constrained Generation. Our approach shares conceptual similarity with inverse problem methods [5,6,18] and other constrained generation approaches [4, 14,41], but with different constraints and task formulations. Overall, HIG adopts a common practice in constrained generation: projecting the unconstrained predictions to the feasible domain. We adapt this idea to the LDM [50] framework and develop an OT-based formulation for enforcing histogram constraints.

Information Embedding. Prior works have focused on hiding data within some medium to avoid detection (also known as steganography). One line of work performs generative image steganography, encoding a secret message (e.g., a bit payload) into a generated image [44,45,58,65,76–78]. A different line instead hides a secret *image* within a coverless container image that is recovered with a key, using text prompts [70], a password-dependent reference image [67], or an image-based semantic feature [66] as the key. In contrast, HIG embeds arbitrary text rather than a fixed bit payload or a secret image, encoding it into the image's color distribution; by leveraging pre-trained LLMs and soft prompts, it remains broadly applicable to any text-to-image model and any form of text.

7 Discussion

Limitations. While our histogram-constrained generation framework enables precise distributional control, several potential improvements are available: 1) For simplicity of implementation, we apply OT-based guidance transformations by randomly sampling pixels or tokens from each source bin and setting them to the target bin values. Better transport strategies are available, such as taking spatial or structural locality into account during sampling, which may reduce the visual artifacts after OT-based transformation; 2) While explicit transformation offers strong control guarantees, it may potentially introduce sharp transitions or inconsistencies. This motivates further exploration of softer designs, such as interpolating between pre- and post-perturbation representations.

Broader Impact. Our method shares the common advantages (e.g., creative content generation) and drawbacks (e.g., harmful or misleading outputs, copyright concerns) of other generative techniques. Beyond these, our information embedding technique enables encoding arbitrary text within visually indistinguishable images, making the hidden content resistant to detection or decoding without access to the exact decoder LLM. While this opens opportunities for secure data transmission, it also raises risks of misuse. A potential safeguard is to train classification models to identify images that are likely to contain embedded information, thus providing an additional layer of defense.

8 Conclusion

We present histogram-constrained image generation (HIG), a novel framework that enables precise, interpretable control in diffusion models by enforcing distributional constraints over pixels or latent tokens. HIG applies explicit OT-based guidance transformations during the denoising process, guiding the generation towards an arbitrary target histogram. Through extensive experiments, we demonstrate the versatility of HIG, including color distribution control, high-capacity information embedding, and latent histogram matching. Our results show that HIG achieves strong alignment with target distributions while preserving visual fidelity, offering a composable, orthogonal control axis that complements existing approaches. Built on a domain-agnostic OT formulation, HIG provides a principled foundation for distributional control that applies wherever the target property is naturally histogram-structured, opening new directions for hybrid and fine-grained controllable synthesis.

Acknowledgement

This work is supported by NYU Shanghai Center for Data Science. This work is supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

1. Balinski, M.L.: Fixed-cost transportation problems. Naval Research Logistics Quarterly **8**(1), 41–54 (1961) 6
2. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 843–852 (2023) 14
3. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation (2024) 13
4. Christopher, J.K., Baek, S., Fioretto, F.: Constrained synthesis with projected diffusion models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=FsdB3I9Y24 14
5. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=OnD9zGAGT0k 14
6. Chung, H., Sim, B., Ryu, D., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=nJJjv0JDJju 14
7. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8795–8805 (2024) 8, 13
8. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013) 6
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 1
10. Dou, Z., Song, Y.: Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=tplXNcHZs1 14
11. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 7, 8
12. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024) 1, 13
13. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 3, 12
14. Fishman, N., Klarner, L., Bortoli, V.D., Mathieu, E., Hutchinson, M.J.: Diffusion models for constrained domains. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=xuWTFQ4VGO, expert Certification 14
15. Gao, J., Liu, Y., Sun, Y., Tang, Y., Zeng, Y., Chen, K., Zhao, C.: Styleshot: A snapshot on any style. arXiv preprint arXiv:2407.01414 (2024) 8, 13
16. Gao, Y., Gong, L., Guo, Q., Hou, X., Lai, Z., Li, F., Li, L., Lian, X., Liao, C., Liu, L., et al.: Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346 (2025) 13

17. Guo, Y., Yang, Y., Yuan, H., Wang, M.: Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models. Advances in Neural Information Processing Systems **38**, 73343–73384 (2026) 14
18. He, Y., Murata, N., Lai, C.H., Takida, Y., Uesaka, T., Kim, D., Liao, W.H., Mitsufuji, Y., Kolter, J.Z., Salakhutdinov, R., Ermon, S.: Manifold preserving guided diffusion. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=o3BxOLoxm1 8, 14
19. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 9
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1, 3
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022) 1, 2, 13
22. Huang, Y., Ghatare, A., Liu, Y., Hu, Z., Zhang, Q., Sastry, C.S., Gururani, S., Oore, S., Yue, Y.: Symbolic music generation with non-differentiable rule guided diffusion. In: Proceedings of the 41st International Conference on Machine Learning. pp. 19772–19797 (2024) 14
23. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022) 13
24. Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., Laine, S.: Analyzing and improving the training dynamics of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24174–24184 (2024) 13
25. Ke, Z., Liu, Y., Zhu, L., Zhao, N., Lau, R.W.: Neural preset for color style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14173–14182 (2023) 13
26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3
27. Labs, B.F.: Flux (2023), https://github.com/black-forest-labs/flux 1, 8, 13, 24
28. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025) 13
29. Larchenko, M., Lobashev, A., Guskov, D., Palyulin, V.V.: Color transfer with modulated flows. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4464–4472 (2025) 13
30. Laria, H., Gomez-Villa, A., Qin, J., Butt, M.A., Raducanu, B., Vazquez-Corral, J., van de Weijer, J., Wang, K.: Leveraging semantic attribute binding for free-lunch color control in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7689–7698 (2026) 13
31. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021) 3, 7, 10

32. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet++: Improving conditional controls with efficient consistency feedback. In: European Conference on Computer Vision. pp. 129–147. Springer (2025) 1, 2, 8, 13, 22
33. Li, X., Zhao, Y., Wang, C., Scalia, G., Eraslan, G., Nair, S., Biancalani, T., Ji, S., Regev, A., Levine, S., et al.: Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. arXiv preprint arXiv:2408.08252 (2024) 14
34. Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C.: Control color: Multimodal diffusion-based interactive image colorization. arXiv preprint arXiv:2402.10855 (2024) 13
35. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t 3, 8, 13
36. Liu, C., Shah, V., Cui, A., Lazebnik, S.: Unziplora: Separating content and style from a single image. arXiv preprint arXiv:2412.04465 (2024) 13
37. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z 8, 13
38. Lobashev, A., Larchenko, M., Guskov, D.: Color conditional generation with sliced wasserstein guidance. Advances in Neural Information Processing Systems **38**, 164572–164601 (2026) 13
39. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019) 8
40. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4296–4304 (2024) 13
41. Naderiparizi, S., Liang, X., Zwartsenberg, B., Wood, F.: Don't be so negative! score-based generative modeling with oracle-assisted guidance (2024), https://openreview.net/forum?id=gJ7cHBHfBk 14
42. OpenAI: Introducing 4o image generation (2025), https://openai.com/index/introducing-4o-image-generation/, accessed: 2025-05-15 8, 13
43. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022) 13
44. Peng, Y., Hu, D., Wang, Y., Chen, K., Pei, G., Zhang, W.: Stegaddpm: Generative image steganography based on denoising diffusion probabilistic model. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 7143–7151. MM '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3612514, https://doi.org/10.1145/3581783.3612514 14
45. Peng, Y., Wang, Y., Hu, D., Chen, K., Rong, X., Zhang, W.: LDStega: Practical and robust generative image steganography based on latent diffusion models. In: ACM Multimedia 2024 (2024), https://openreview.net/forum?id=kEqGgMgIlu 14
46. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 1, 3, 8, 13, 23
47. Qiu, Q., Mao, J., Wang, X.: Exploring palette based color guidance in diffusion models. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10287–10295 (2025) 13

48. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. arXiv preprint arXiv:2412.03069 (2024) 3, 12
49. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019) 3
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 3, 13, 14
51. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (June 2023) 1, 2, 8, 13, 22
52. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 9
53. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. In: European Conference on Computer Vision. pp. 422–438. Springer (2025) 13
54. Shum, K.C., Hua, B.S., Nguyen, D.T., Yeung, S.K.: Color alignment in diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28446–28455 (2025) 13
55. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 1
56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP 1, 3, 8
57. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) 3
58. Su, W., Ni, J., Sun, Y.: Stegastylegan: Towards generic and practical generative image steganography. Proceedings of the AAAI Conference on Artificial Intelligence **38**(1), 240–248 (Mar 2024). https://doi.org/10.1609/aaai.v38i1.27776 14
59. Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025) 13
60. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 3, 4
61. Villani, C.: Optimal Transport: Old and New. Springer (2009) 2, 4
62. Wang, H., Xing, P., Huang, R., Ai, H., Wang, Q., Bai, X.: Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. arXiv preprint arXiv:2407.00788 (2024) 8, 13
63. Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., Yang, J.: Seededit 3.0: Fast and high-quality generative image editing. arXiv preprint arXiv:2506.05083 (2025) 13
64. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S.: Sana: Efficient high-resolution image synthesis with linear diffusion transformer (2024) 13

65. Xu, Z., xu, D., Li, Z., Zhang, C.: MDDM: Practical message-driven generative image steganography based on diffusion models. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=NniXePXVXw 14
66. Yan, L., Li, X., Zhang, J., Guan, F., Peng, K., Li, P.: F-ddim: A featurized denoising diffusion implicit model for facial image steganography. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 8488–8496. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755517, https://doi.org/10.1145/3746027.3755517 14
67. Yang, Y., Liu, Z., Jia, J., Gao, Z., Li, Y., Sun, W., Liu, X., Zhai, G.: Diffstega: towards universal training-free coverless image steganography with diffusion models. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 1579–1587 (2024) 14
68. Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J.Y., Ermon, S.: Tfg: Unified training-free guidance for diffusion models. Advances in Neural Information Processing Systems **37**, 22370–22417 (2024) 14
69. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023) 8
70. Yu, J., Zhang, X., Xu, Y., Zhang, J.: Cross: Diffusion model makes controllable, robust and secure image steganography. Advances in Neural Information Processing Systems **36**, 80730–80743 (2023) 14
71. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. Advances in Neural Information Processing Systems **37**, 128940–128966 (2024) 3, 12
72. Zamzam, O.: Pixelshuffler: A simple image translation through pixel rearrangement. arXiv preprint arXiv:2410.03021 (2024) 8, 13
73. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1, 2, 13
74. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10146–10156 (June 2023) 8, 13
75. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024) 13
76. Zhou, Q., Wei, P., Qian, Z., Zhang, X., Li, S.: Improved generative steganography based on diffusion model. IEEE Transactions on Circuits and Systems for Video Technology **35**(7), 6494–6507 (2025). https://doi.org/10.1109/TCSVT.2025.3539832 14
77. Zhou, Z., Dong, X., Meng, R., Wang, M., Yan, H., Yu, K., Choo, K.K.R.: Generative steganography via auto-generation of semantic object contours. IEEE Transactions on Information Forensics and Security **18**, 2751–2765 (2023). https://doi.org/10.1109/TIFS.2023.3268843 14
78. Zhou, Z., Su, Y., Li, J., Yu, K., Wu, Q.M.J., Fu, Z., Shi, Y.: Secret-to-Image Reversible Transformation for Generative Steganography . IEEE Transactions on Dependable and Secure Computing **20**(05), 4118–4134 (Sep 2023). https://doi.org/10.1109/TDSC.2022.3217661 14